A global overview of pleiotropy and genetic architecture in complex traits

Watanabe, Kyoko; Stringer, Sven; Frei, Oleksandr; Umievi Mirkov, Maša; de Leeuw, Christiaan; Polderman, Tinca J C; van der Sluis, Sophie; Andreassen, Ole A; Neale, Benjamin M; Posthuma, Danielle

published in
Nature Genetics
2019

DOI (link to publisher)
10.1038/s41588-019-0481-0

document version
Publisher's PDF, also known as Version of record

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0481-0

Corrected: Author Correction

A global overview of pleiotropy and genetic architecture in complex traits

Kyoko Watanabe[1], Sven Stringer[1], Oleksandr Frei[2], Maša Umićević Mirkov[1], Christiaan de Leeuw[1], Tinca J. C. Polderman[1], Sophie van der Sluis[1,3], Ole A. Andreassen[2,4], Benjamin M. Neale[5,6,7] and Danielle Posthuma[1,3*]

After a decade of genome-wide association studies (GWASs), fundamental questions in human genetics, such as the extent of pleiotropy across the genome and variation in genetic architecture across traits, are still unanswered. The current availability of hundreds of GWASs provides a unique opportunity to address these questions. We systematically analyzed 4,155 publicly available GWASs. For a subset of well-powered GWASs on 558 traits, we provide an extensive overview of pleiotropy and genetic architecture. We show that trait-associated loci cover more than half of the genome, and 90% of these overlap with loci from multiple traits. We find that potential causal variants are enriched in coding and flanking regions, as well as in regulatory elements, and show variation in polygenicity and discoverability of traits. Our results provide insights into how genetic variation contributes to trait variation. All GWAS results can be queried and visualized at the GWAS ATLAS resource (https://atlas.ctglab.nl).

Since the first genome-wide association study on macular degeneration in 2005 (ref. [1]), over 3,000 GWASs have been published, for over 1,000 traits, reporting on tens of thousands of genetic risk variants[2]. These results have increased our understanding of the genetic architecture of traits. Occasionally, GWAS results have led to further insight into disease mechanisms[3,4], such as autophagy for Crohn's disease[5], immunodeficiency for rheumatoid arthritis[6] and transcriptome regulation through *FOXA2* in the pancreatic islet and liver for type 2 diabetes[7]. After a decade of GWASs, we have learned that the majority of studied traits are highly polygenic and influenced by many genetic variants, each of small effect[4,8], with disparate genetic architectures across traits[9]. Fundamental questions (such as whether all genetic variants or genes in the human genome are associated with at least one, many or even all traits, and whether the polygenic effects for specific traits are functionally clustered or randomly spread across the genome) are, however, still unanswered[4,10,11]. Such knowledge would greatly enhance our understanding of how genetic variation leads to trait variation and trait correlations. Whereas GWAS primarily aims to discover genetic variants associated with specific traits, the current availability of vast amounts of GWAS results allow investigation of these general questions.

To this end, we compiled a catalog of 4,155 GWAS results across 2,965 unique traits from 295 studies (https://atlas.ctglab.nl), including publicly available GWASs and new results for 600 traits from the UK Biobank[12]. These GWAS results were used in the current study to (1) chart the extent of pleiotropy at trait-associated locus, gene, SNP and gene-set levels, (2) characterize the nature of trait-associated variants (that is, the distribution of effect size, minor allele frequency (MAF) and biological functionality of trait-associated or credible SNPs) and (3) investigate genetic architecture across a variety of traits and domains in terms of SNP heritability and trait polygenicity (see Supplementary Fig. 1).

Results

Catalog of 4,155 GWAS summary statistics. We collected publicly available, full GWAS summary statistics (last update 23 October 2018; see Methods) resulting in 3,555 sets of GWAS summary statistics from 294 studies. We additionally performed GWAS on 600 traits available from the UK Biobank release 2 cohort (UKB2; release May 2017)[12], by selecting nonbinary traits with >50,000 European individuals with nonmissing phenotypes, and binary traits for which the number of available cases and controls were both >10,000 and total sample size was >50,000 (see Methods, Supplementary Note and Supplementary Tables 1 and 2). In total, we collected 4,155 GWASs from 295 unique studies covering 2,965 unique traits (Supplementary Table 3). Traits were classified into 27 domains[13,14]. The average sample size across curated GWASs was 56,250 subjects, with a maximum of 898,130 for type 2 diabetes[15]. The 4,155 GWAS results are made available in an online database (https://atlas.ctglab.nl), which provides a variety of information per trait, including SNP-based and gene-based Manhattan plots, gene-set analyses[16], SNP heritability estimates[17], genetic correlations, cross-GWAS comparisons and phenome-wide plots.

We restricted subsequent analyses to reasonably powered GWASs ($N > 50{,}000$), to avoid including effect estimates with relatively large standard errors (see Methods). For each unique trait, we selected the GWAS with the largest sample size, resulting in 558 GWASs for 558 unique traits across 24 trait domains (479 GWASs based on UKB2, Supplementary Table 3). All results presented

[1]Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University Amsterdam, Amsterdam, the Netherlands. [2]NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [3]Department of Clinical Genetics, Section of Complex Trait Genetics, Neuroscience Campus Amsterdam, VU Medical Center, Amsterdam, the Netherlands. [4]Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway. [5]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [6]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [7]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. *e-mail: d.posthuma@vu.nl

Table 1 | Count and proportion of pleiotropic trait-associated loci, genes, SNPs and gene sets

	Loci		Genes		SNPs		Gene set	
	Length (Mb)	%	Count	%	Count	%	Count	%
Total in genome	2,796.10	100.00	17,518	100.00	1,740,179	100.00	10,086	100.00
Associated	1,706.97	61.05	11,544	65.90	236,638	13.60	1,030	10.21
Pleiotropic[a]	1,593.32	93.34	9,374	81.20	142,439	60.19	587	56.99
Multidomain	1,537.06	90.04	7,754	67.17	76,703	32.41	353	34.27
Domain specific	56.26	3.30	1,620	14.03	65,736	27.78	234	22.72
Trait specific	113.64	6.66	2,170	18.80	94,199	39.81	443	43.01
Nonassociated	1,089.13	38.95	5,974	34.10	1,503,541	86.40	9,056	89.79

[a]The count of pleiotropic loci, genes, SNPs and gene sets is the sum of the multidomain and domain-specific categories. Proportion of pleiotropic, multidomain, domain-specific and trait-specific categories are relative to the associated loci, SNPs, genes or gene sets, respectively.

hereafter concern these selected 558 GWASs unless otherwise specified. The online database, however, allows researchers to reproduce similar analyses with custom selections of GWASs.

Extent of pleiotropy. Results of previous GWASs showed significant associations of thousands of genomic loci with a large number of traits[2,4]. Given a finite number of segregating variants on the human genome, this suggests the presence of widespread pleiotropy. Pleiotropy may inform reasons for comorbidity between traits, pointing to underlying shared genetic mechanisms, and may aid in establishing the direction of causality between traits. Currently, the exact extent of pleiotropy across the genome is unknown[4]. We therefore aimed to quantify the extent of pleiotropy. We defined pleiotropy as the presence of statistically significant associations with more than one trait domain, as traits within a domain tend to show stronger phenotypic correlations than those between domains (see Supplementary Note and Supplementary Fig. 2). Our definition thus refers to 'statistical pleiotropy', and includes situations of true pleiotropy (for example, one SNP directly influences multiple traits, or different causal variants are present for two traits, but these are in high linkage disequilibrium (LD)), and situations where statistical associations to multiple traits are induced via causal effects of one trait on another, via phenotypic correlations between traits or via a third common factor[18]. The level of pleiotropy was grouped into three categories: multidomain (associated with traits from multiple domains), domain specific (associated with multiple traits from a single domain) and trait specific (associated with a single trait; see Methods). We then assessed whether pleiotropic associations at the locus, gene, SNP or gene-set level are structurally or functionally different from nonassociated sites.

Pleiotropic genomic loci. The 558 GWASs yielded 41,533 trait-associated loci (from 470 traits; 88 traits did not yield any genome-wide significant associations; see Methods). Grouping physically overlapping trait-associated loci resulted in 3,362 loci (see Methods, Supplementary Fig. 3 and Supplementary Table 4), with a summed length of 1,707.0 megabases (Mb) covering 61.0% of the genome. Of these, 93.3% were loci associated with more than one trait, and 90.0% were multidomain loci (see Table 1 and Supplementary Fig. 4a,b). The multidomain and domain-specific loci showed a significantly higher density of protein-coding genes compared with nonassociated genomic regions ($P=4.4\times10^{-16}$ and $P=3.7\times10^{-4}$, two-sided Mann–Whitney *U*-test; Fig. 1a and Supplementary Table 5). Additionally, trait-associated loci are more densely overlapping with loci from other traits than expected under the null hypothesis of no pleiotropy (Supplementary Note).

The most pleiotropic locus associated with the largest number of traits and domains was the MHC region (chr6: 25–37 Mb), containing 441 trait-associated loci from 213 traits across 23 trait domains. The MHC region is well known for its high degree of LD, spanning over 300 genes. The extremely pleiotropic nature of this region is thus partly explained by its long-ranged LD blocks and overlap of multiple independent signals from multiple traits. High locus pleiotropy, not limited to the MHC region, can occur purely due to the overlap of the LD blocks of the loci in a grouped locus, and they may not share the same causal SNPs. By performing colocalization (that is, statistically identifying loci sharing the same causal SNP) for all possible pairs of physically overlapping trait-associated loci (see Methods and Supplementary Fig. 3), 35,609 loci (88.4% of 40,262 loci physically overlapping with at least one locus from other traits) were colocalized with at least one other locus, and 22,319 loci (55.4%) colocalized with at least one locus of a trait from a different trait domain (see Supplementary Note). We indeed observed an average decrease of 38.3% in the number of associated trait domains per group of colocalized loci compared to grouped loci defined by physical overlap (see Supplementary Note, Supplementary Fig. 4 and Supplementary Table 6). In addition, loci grouped based on physical overlap often contained multiple independent groups of colocalized loci (Supplementary Table 6). Therefore, physical overlap of trait-associated loci does not necessarily mean that the same causal SNPs are involved in the traits associated with such a grouped locus. Examination of pleiotropy at the gene or SNP level provides further insight into the nature of the pleiotropy.

Pleiotropic genes. To investigate the extent of pleiotropy at the gene level, we conducted gene-based analyses for each trait on 17,518 protein-coding genes using MAGMA[16] (see Methods). Of the 558 traits, 518 yielded at least one associated gene, and 11,544 (65.9%) genes were associated with at least one trait (Supplementary Table 7). Of these, 81.2% were associated with more than one trait and 67.2% with traits from multiple domains (see Table 1 and Supplementary Fig. 5a,b). We found that genes associated with at least one trait are significantly longer than genes not associated with any of the 558 tested traits ($P=2.3\times10^{-192}$, $P=6.9\times10^{-12}$ and $P=5.1\times10^{-29}$ for multidomain, domain-specific and trait-specific genes, respectively, two-sided *t*-test; Fig. 1b and Supplementary Table 8). As the MAGMA algorithm accounts for gene length, these findings are unlikely to be due to larger genes having an increased statistical probability to be significantly associated (see Supplementary Note, Supplementary Fig. 5c and Supplementary Table 9). The multidomain genes showed a significantly higher probability of being intolerant to loss of function mutations (pLI score)[19] compared with trait-, domain-specific and nonassociated genes ($P=2.2\times10^{-81}$, $P=3.2\times10^{-22}$ and $P=1.5\times10^{-18}$, respectively, two-sided Mann–Whitney *U*-test; Fig. 1c and Supplementary Table 10). The most pleiotropic genes are located in the MHC region, yet, a region on chromosome 3 also spanned multiple genes with high levels of pleiotropy (Supplementary Fig. 5a).

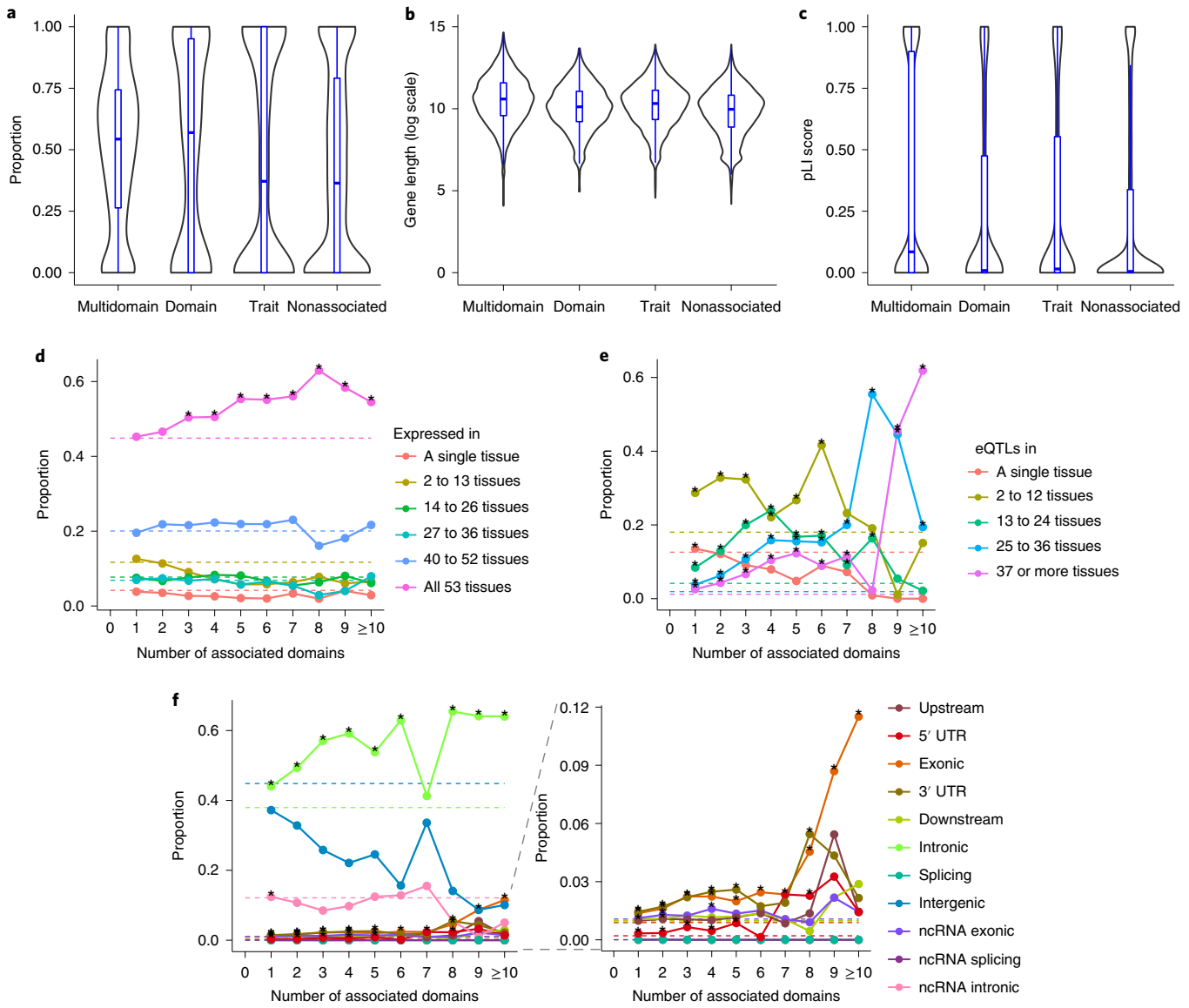

Fig. 1 | Trait-associated locus, gene and SNP pleiotropy across the genome. a, Distribution of gene density of loci with different association types. **b**, Distribution of gene length in log scale with different association types. **c**, Distribution of pLI scores of genes with different association types. **a**–**c**, Multidomain: associated with traits from >1 domain, domain: associated with >1 trait from a single domain, trait: associated with a single trait, nonassociated: not associated with any of 558 traits. **d**, Tissue specificity of genes at different levels of pleiotropy. Each data point represents a proportion of genes expressed in a given number of tissues for a specific number of associated domains. **e**, Tissue specificity of SNPs based on active eQTLs at different levels of pleiotropy. Each data point represents the proportion of SNPs being eQTLs in a given number of tissues for a specific number of associated domains. **f**, Proportion of SNPs with different functional consequences at different levels of pleiotropy. Each data point represents the proportion of SNPs with a given functional consequence for a specific number of associated domains. **d**–**f**, Dashed lines refer to the baseline proportions (relative to all 17,444 genes (**d**) or all 1,740,179 SNPs (**e**,**f**)) and stars denote significant enrichment relative to the baseline (one-sided Fisher's exact test).

We next tested whether the tissue specificity of genes was related to the level of pleiotropy by counting the number of active tissues per gene based on gene expression profiles for 53 tissues obtained from GTEx[20] (see Methods). Indeed, the proportion of genes expressed in all 53 tissues increases along with the level of pleiotropy ($P=4.7\times10^{-4}$ for regression coefficient; Fig. 1d and Supplementary Table 11), indicating that more pleiotropic genes tend to be active in multiple tissue types, and suggesting that those genes are involved in general biological functions across the human body.

Pleiotropic SNPs. Within the same locus or gene, multiple SNPs may be significantly associated with different traits. A locus or gene can thus show higher levels of pleiotropy than individual SNPs. To investigate the extent of pleiotropy at the level of SNPs, we extracted 1,740,179 SNPs present in all 558 GWASs. We confirmed that this subset of SNPs was not strongly structurally biased in terms of genome coverage ($r=0.98$, $P=0.02$ with null hypothesis of $r=1$) and functional consequences ($r=1.00$, $P=0.07$) compared with all known SNPs on the genome (see Methods and Supplementary Fig. 6a,b). Of the 1.7 million SNPs analyzed, 236,638 (13.6%) were genome-wide significant ($P<5\times10^{-8}$) in at least one trait (Supplementary Fig. 6c and Supplementary Table 12). Of these, 60.2% were associated with more than one trait and 32.4% showed multidomain associations (Table 1 and Supplementary Fig. 6d).

These pleiotropic SNPs were spread widely across the genome but were not evenly distributed, with chromosomes 1, 5, 11, 12,

15, 17, 20 and 22 showing relative enrichment of pleiotropic SNPs (see Supplementary Note and Supplementary Table 13). Of all trait-associated SNPs, the most pleiotropic SNP, located in the MHC region (rs707939; an intronic SNP of *MSH5*) was associated with 48 traits from 13 domains. There were 41 SNPs associated with 12 trait domains, of which 35 were located on chromosome 3, 49.8–50.1 Mb overlapping with five protein-coding genes, *TRAIP*, *CAMKV*, *MST1R*, *MON1A* and *RBM6*. These SNPs include two exonic SNPs, on *CAMKV* (synonymous) and *MST1R* (nonsynonymous; Supplementary Table 12).

To investigate whether more pleiotropic SNPs are functionally different from less pleiotropic SNPs, we investigated how functional consequence and tissue specificity, in terms of expression quantitative trait loci (eQTLs based on GTEx), were represented across different levels of SNP pleiotropy (see Methods). With increasing levels of pleiotropy, the proportion of exonic SNPs increased from less than 1% to over 5% ($P=1.6\times10^{-2}$ for regression coefficient), and the proportion of intronic SNPs increased from less than 40% to over 50% ($P=2.2\times10^{-3}$; Fig. 1e and Supplementary Table 14). The proportion of SNPs within flanking regions such as 5′ and 3′ untranslated regions (UTRs) also increased with the number of associated domains. Concurrently, we observed a steep decrease in the proportion of intergenic SNPs with increasing level of SNP pleiotropy ($P=6.8\times10^{-4}$; Fig. 1e and Supplementary Table 14). Based on active eQTLs, the proportion of SNPs being eQTLs in a greater number of tissues (>24 tissues out of 48) increased, along with the number of associated domains ($P=7.4\times10^{-3}$ and $P=1.1\times10^{-2}$ for eQTLs in between 25 and 36 tissues, and between 37 and 48 tissues, respectively) while SNPs in genes expressed in a single, or in less than half of, available tissue types showed a decreasing proportion (Fig. 1f and Supplementary Table 15). These results suggest that highly pleiotropic SNPs are more likely to be genic (exonic or intronic), and less likely to be tissue specific.

Pleiotropic gene sets. Pleiotropy at the level of trait-associated loci, genes or SNPs does not necessarily suggest the presence of shared biological pathways across multiple traits. To assess the level of pleiotropy at the functional level, we performed gene-set analyses for the 558 traits using 10,086 gene sets (see Methods). In total, 218 (39.1%) traits showed significant associations with one of the 1,030 (10.2%) gene sets. The most pleiotropic gene set was 'Regulation of transcription from RNA polymerase II promoter', associated with 74 traits from 10 domains, followed by 7 gene sets associated with ≥7 domains, including 5 gene sets involved in regulation of transcription (Supplementary Table 16). The number of genes in a gene set was significantly larger for highly pleiotropic gene sets (multidomain) than for other gene sets ($P=6.8\times10^{-10}$, $P=3.3\times10^{-17}$ and $P=6.9\times10^{-33}$ for domain specific, trait specific and nonassociated, respectively, two-sided *t*-test; Supplementary Fig. 7a and Supplementary Table 17).

In contrast to the gene pleiotropy where 80.9% of genes were associated with more than one trait, only 57.0% of the associated gene sets were pleiotropic (Table 1). Additionally, the proportion of pleiotropic genes per gene set is not uniformly distributed, and pleiotropic genes tend to cluster into a subset of gene sets, explaining the decreased proportion of pleiotropic gene sets compared with pleiotropic genes (Supplementary Note and Supplementary Fig. 7b,c). At the same time, the higher proportion of trait-specific gene sets (43.0%) compared with trait-specific genes (18.8%) suggests that, given currently defined gene sets, the combination of associated genes is rather specific to each trait.

Genetic correlations across traits. Above we showed that of all trait-associated loci, genes and SNPs that are associated with at least one trait, 90.0%, 67.2% and 32.4%, respectively, are associated with traits from multiple domains. Such widespread pleiotropy indicates nonzero genetic correlations between traits. To test whether genetic correlations are evenly distributed across traits or cluster into trait domains, we computed pairwise genetic correlations (r_g) across 558 traits using LD score regression (LDSC)[17].

We calculated the proportion of trait pairs with an r_g significantly different from zero across all 558 traits, within and between domains. Out of 155,403 possible pairs across 558 traits (average $|r_g|$ of 0.16), 24,170 pairs (15.5%) showed significant genetic correlations after Bonferroni correction ($P<0.05/155{,}403=3.2\times10^{-7}$) with an average $|r_g|$ of 0.38.

If the trait domains contain traits that are biologically related, we would expect traits within the same domain to have stronger genetic correlations than traits across domains. Indeed, most of the domains showed a proportion of trait pairs with significant genetic correlations of >20% and an average $|r_g|>0.5$ within the domains (Fig. 2a and Supplementary Table 18). The proportion of pairs with a significant genetic correlation within domains was especially high in cognitive, 'ear, nose, throat', metabolic and respiratory domains. Note that the proportion of trait pairs with significant r_g may be biased by sample size and SNP heritability (h^2_{SNP}) of traits within a domain; across 558 traits, the worst-case scenarios with the minimum observed h^2_{SNP} (0.0045 with sample size 385,289) or the minimum sample size (51,750 with $h^2_{SNP}=0.0704$) required r_g to be above 0.39 or 0.18, respectively, to gain a power of 0.8 (see Methods). The proportion of trait pairs with significant genetic correlations was generally lower between domains than within domains, and most of the domain pairs showed average $|r_g|<0.4$ (Fig. 2b and Supplementary Table 19). We further clustered traits based on genetic correlations (using $|r_g|$), which resulted in the majority of clusters containing traits from multiple domains (see Methods, Supplementary Note and Supplementary Fig. 8). These results suggest that, although $|r_g|$ is higher within domains than across domains, the current definition of trait domains does not necessarily comprise genetically similar traits.

Nature of trait-associated variants. We investigated characteristics of trait-associated variants in terms of their effect sizes, MAF, functional consequences on genes and regulatory functions. We extracted all lead SNPs from each of the 558 GWASs. Lead SNPs were defined per trait at the standard threshold for genome-wide significance ($P<5\times10^{-8}$) and using an r^2 of 0.1 to obtain near-independent lead SNPs, based on the population-relevant reference panel (see Methods). This resulted in 82,633 lead SNPs for 476 traits, reflecting 43,492 unique SNPs. Out of 558 traits, 82 traits did not yield any genome-wide significant lead SNP after quality control.

Distribution of MAF and effect sizes of lead SNPs. Of the 43,492 (unique) lead SNPs derived from the 558 GWASs, 12.3% had a MAF below 0.01, which is significantly less than expected given the proportion of rare variants in the reference panels ($P<1\times10^{-323}$, two-sided Fisher's exact test; Supplementary Note), while the distribution of lead SNPs with a MAF above 0.01 was nearly uniform (Fig. 3a).

We calculated the standardized effect size (β) from *Z*-statistics as a function of MAF and sample size[21], and inspected the distribution of the squared standardized effect sizes (β^2) for lead SNPs across all traits (see Methods). The median β^2 of the lead SNPs across all traits was 5.7×10^{-4} (4.9×10^{-4} and 6.0×10^{-2} for lead SNPs with MAF ≥ 0.01 and <0.01, respectively), and 94.6% of lead SNPs had $\beta^2<0.05$ (Fig. 3b). We observed a relationship between MAF and β^2, with rare variants (MAF < 0.01) showing larger effect sizes (Fig. 3c), corresponding with the notion that rare variants are more likely to have large effects than common variants, as they are less likely to be under strong selective pressure[22]. However, we also note that statistical power for detecting rare variants is unstable[23]. Given that the proportion of rare lead SNPs is larger than other MAF bins,

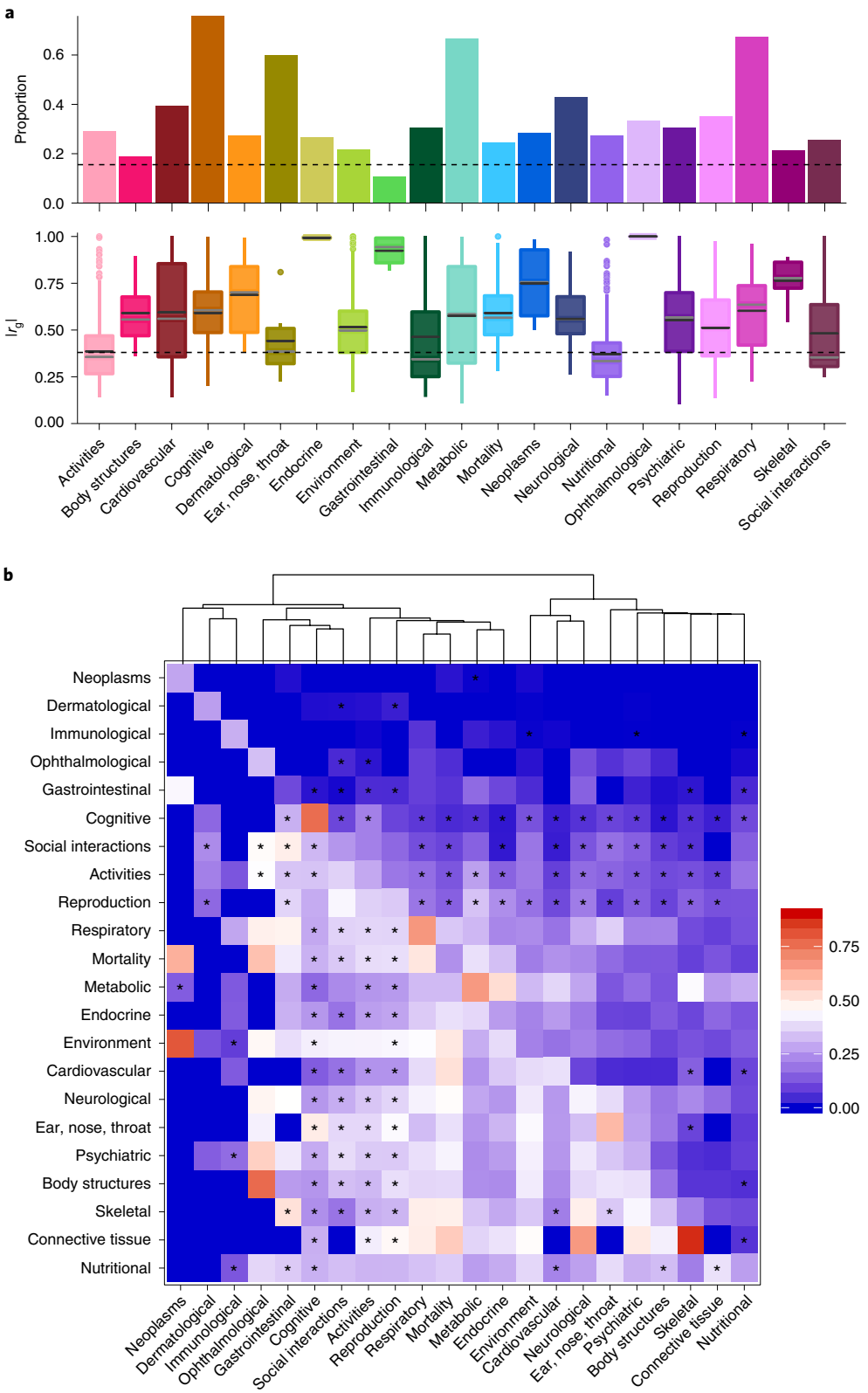

Fig. 2 | Within- and between-domain genetic correlations. a, Proportion of trait pairs with significant r_g (top) and distribution absolute values of genetic correlation ($|r_g|$) for significant trait pairs (bottom) within domains. Dashed lines represent the proportion of trait pairs with significant r_g (top) and average $|r_g|$ for significant trait pairs (bottom) across all 558 traits, respectively. In the box plots (bottom panel), dark gray and light gray horizontal lines represent mean and median, respectively. Connective tissue, muscular and infection domains are excluded, as these each contains less than three traits. **b**, Heat map of proportion of trait pairs with significant r_g (upper right triangle) and average $|r_g|$ for significant trait pairs (lower left triangle) between domains. Connective tissue, muscular and infection domains are excluded, as each contains less than three traits. The diagonal represents the proportion of trait pairs with significant r_g within domains. Stars denote the pairs of domains in which the majority (>50%) of significant r_g are negative.

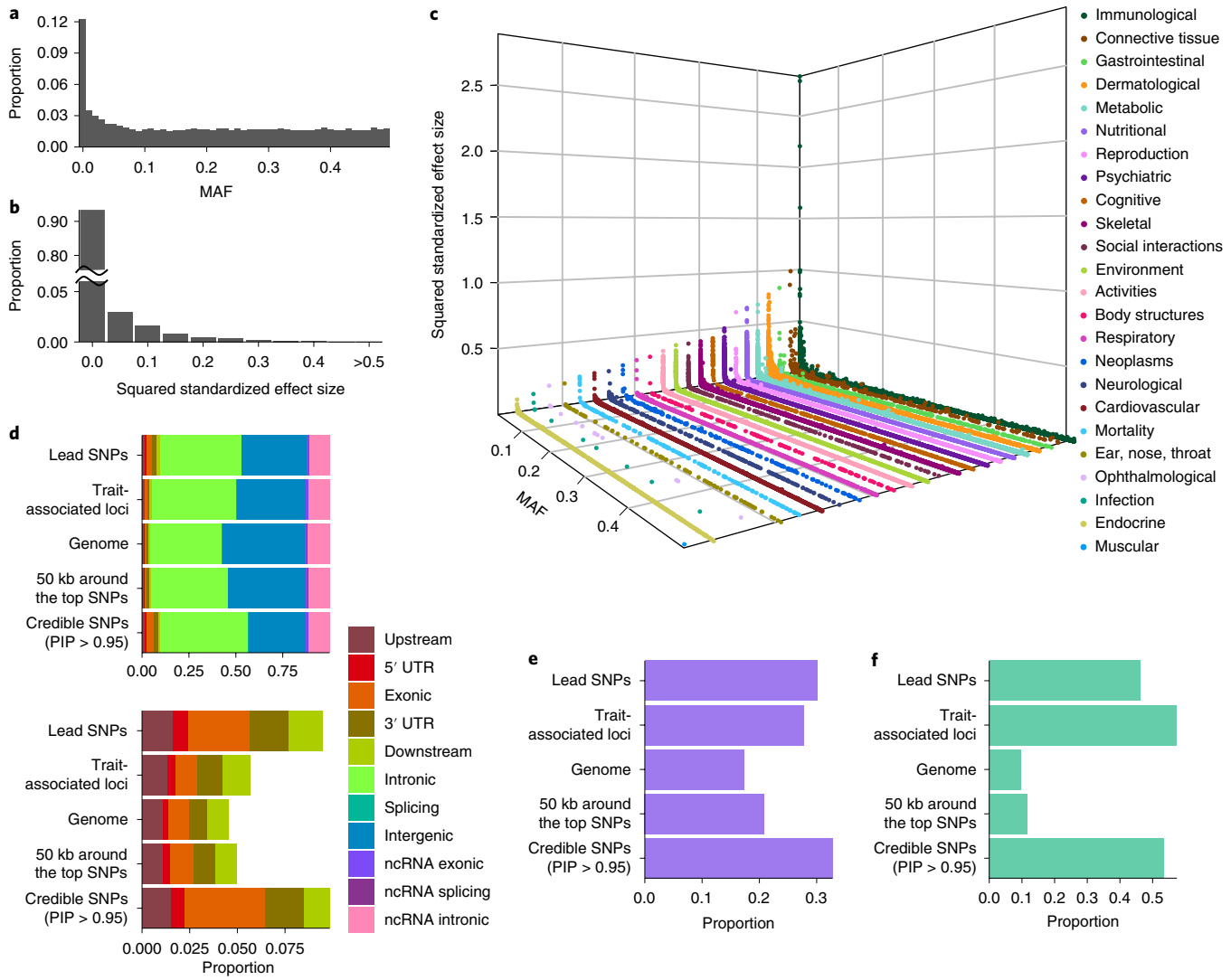

Fig. 3 | Distribution and characterization of lead SNPs and credible SNPs of 558 traits. a, Histogram of MAF of the unique lead SNPs. **b**, Histogram of squared standardized effect size of lead SNPs. **c**, Scatter plot of MAF and squared standardized effect sizes of lead SNPs grouped by trait domains. **d**, Distribution of functional consequences of SNPs. **e**, Proportion of SNPs that overlap with active consequence chromatin state (≤7) across 127 tissue/cell types. **f**, Proportion of SNPs overlapping with significant eQTLs from any of 48 available tissue types.

the distribution of the effect sizes may have longer tails for SNPs with MAF < 0.01. For most traits, a similar relationship between MAF and standardized effect size was observed (Supplementary Fig. 9), but large variation across traits was seen in terms of the number of rare lead SNPs, with, for example, a large proportion of rare variants influencing nutritional and social interaction domains (possibly due to the larger sample sizes; Supplementary Note, Supplementary Fig. 10 and Supplementary Tables 20 and 21).

Characterization of trait-associated loci and lead SNPs. We sought to characterize differences in the distribution of functional annotations comparing SNPs within trait-associated loci to all SNPs in the genome, and comparing lead SNPs to SNPs in the trait-association loci (see Methods). Comparing SNPs in the trait-associated loci against the entire genome, the strongest enrichment of SNPs in trait-associated loci was seen in flanking regions (upstream, downstream, 5′ and 3′ UTR) with average fold enrichment (E) of 1.31 (Fig. 3d and Table 2). Of SNPs in trait-associated loci, 93.1% were noncoding, where intergenic SNPs were significantly depleted ($E = 0.84$), while intronic SNPs were significantly enriched compared with all SNPs in the genome ($E = 1.17$; Table 2). SNPs in trait-associated loci were also more often exonic compared to the entire genome ($E = 1.07$). Active chromatin states and eQTLs were also significantly enriched, with notably high enrichment of eQTLs ($E = 1.61$ and 5.95, respectively; Table 2).

We next compared lead SNPs with SNPs in the trait-associated loci. The strongest enrichment was seen in exonic SNPs ($E = 2.84$) followed by flanking regions ($E = 1.38$), while intronic and intergenic regions were depleted (average $E = 0.95$; Fig. 3d and Table 2). These results clearly indicate that SNPs located in exonic and flanking regions tend to show stronger effect sizes than other SNPs within the trait-associated loci. On the other hand, active chromatin states showed significant enrichment with a slight increased proportion ($E = 1.08$) while eQTLs were significantly depleted ($E = 0.80$) compared to SNPs in the trait-associated loci (Fig. 3e,f and Table 2). This suggests that SNPs within the trait-associated loci largely overlap with regulatory elements but that these elements do not always have the strongest effect sizes within the loci.

Characterization of a credible set of SNPs based on fine mapping. Lead SNPs (that is, defined by LD and P values) are not necessarily the causal SNPs in trait-associated loci[24]. We therefore performed fine

Table 2 | Characteristics of lead SNPs and credible SNPs with PIP > 0.95 across 558 traits versus all SNPs in the genome

Annotation categories	Genome	Trait-associated loci			Lead SNPs			SNPs in fine-mapped regions[a]			Credible SNPs (PIP > 0.95)[b]		
	%	%	*E*	*P*[c]	%	*E*	*P*[d]	%	*E*	*P*[c]	%	*E*	*P*[e]
Noncoding	94.37	93.06	0.99	$<1\times10^{-323}$	89.13	0.96	1.54×10^{-201}	93.95	1.00	$<1\times10^{-323}$	89.10	0.95	4.06×10^{-90}
Intergenic	44.11	36.88	0.84	$<1\times10^{-323}$	34.31	0.93	1.85×10^{-29}	41.45	0.94	$<1\times10^{-323}$	30.79	0.74	1.01×10^{-127}
Intronic	38.29	44.88	1.17	$<1\times10^{-323}$	43.85	0.98	1.15×10^{-5}	41.04	1.07	$<1\times10^{-323}$	46.96	1.14	3.82×10^{-39}
scRNA intronic	11.98	11.29	0.94	2.38×10^{-125}	10.98	0.97	3.47×10^{-2}	11.47	0.96	$<1\times10^{-323}$	11.35	0.99	6.98×10^{-1}
Coding	2.15	2.40	1.12	2.56×10^{-79}	4.60	1.92	3.07×10^{-161}	2.31	1.07	$<1\times10^{-323}$	5.29	2.29	1.79×10^{-77}
Exonic	1.06	1.13	1.07	1.64×10^{-15}	3.22	2.84	9.60×10^{-257}	1.22	1.15	$<1\times10^{-323}$	4.24	3.47	5.35×10^{-122}
Splicing	1.16×10^{-2}	1.13×10^{-2}	0.98	0.828532	2.11×10^{-2}	1.86	6.20×10^{-2}	0.01	1.13	3.30×10^{-28}	0.02	1.28	6.72×10^{-1}
ncRNA exonic	1.07	1.25	1.16	2.72×10^{-77}	1.36	1.09	4.55×10^{-2}	1.07	1.00	7.09×10^{-1}	1.04	0.97	7.90×10^{-1}
ncRNA splicing	5.40×10^{-3}	5.09×10^{-3}	0.94	0.702988	2.35×10^{-3}	0.46	7.27×10^{-1}	0.01	0.98	3.14×10^{-1}	0.00	0.00	1
Flanking regions	3.48	4.54	1.31	$<1\times10^{-323}$	6.27	1.38	1.02×10^{-61}	3.74	1.07	$<1\times10^{-323}$	5.60	1.50	9.17×10^{-24}
Upstream	1.09	1.33	1.22	3.83×10^{-135}	1.64	1.23	3.19×10^{-8}	1.12	1.02	1.64×10^{-67}	1.52	1.36	5.96×10^{-5}
5′ UTR	0.30	0.44	1.48	3.44×10^{-166}	0.78	1.76	8.18×10^{-23}	0.35	1.17	$<1\times10^{-323}$	0.70	2.01	6.28×10^{-9}
3′ UTR	0.98	1.32	1.34	2.08×10^{-285}	2.06	1.56	2.63×10^{-36}	1.14	1.16	$<1\times10^{-323}$	2.04	1.79	6.07×10^{-17}
Downstream	1.10	1.45	1.32	1.18×10^{-280}	1.79	1.23	7.82×10^{-9}	1.13	1.03	4.20×10^{-22}	1.34	1.18	3.78×10^{-2}
Active chromatin	17.24	27.74	1.61	$<1\times10^{-323}$	30.10	1.08	8.45×10^{-30}	20.86	1.21	$<1\times10^{-323}$	32.75	1.57	9.05×10^{-202}
eQTLs	9.66	57.41	5.95	$<1\times10^{-323}$	46.15	0.80	$<1\times10^{-323}$	11.47	1.19	$<1\times10^{-323}$	53.56	4.67	$<1\times10^{-323}$

E, fold enrichment (proportion of SNPs with a certain annotation divided by the proportion of SNPs with the same annotation in background). [a]Fine-mapped regions are 50 kb windows from the top SNPs or the trait-associated loci, whichever is larger. [b]From 95% credible set SNPs, only SNPs with PIP > 0.95 were selected. [c]*P* value of Fisher's exact test (two-sided) against the entire genome. [d]*P* value of Fisher's exact test (two-sided) against trait-associated loci. [e]*P* value of Fisher's exact test (two-sided) against 50 kb around the top SNPs.

mapping using FINEMAP[25] for 41,041 trait-associated loci from 466 traits to obtain credible SNPs, and characterized these in the same way as was done for lead SNPs (see Methods, Supplementary Note and Supplementary Fig. 11). The enrichment pattern of SNPs in the fine-mapped regions was similar to SNPs in the trait-associated loci; that is, significant enrichments in exonic, intronic and flanking regions, active chromatin state and eQTLs (Fig. 3d and Table 2; see Supplementary Note for details). Credible SNPs (with posterior inclusion probability (PIP) > 0.95) showed similar enrichment patterns to lead SNPs; strong enrichment in exonic ($E = 3.47$) and flanking regions ($E = 1.50$), as well as intronic regions ($E = 1.14$) compared to the SNPs in fine-mapped regions (Table 2). The credible SNPs were also significantly enriched in both active chromatin states ($E = 1.57$) and eQTLs ($E = 4.67$; Fig. 3e,f and Table 2). Notably, enrichments of credible SNPs in exonic and flanking regions, active chromatin and eQTLs were much higher in credible SNPs compared with lead SNPs, indicating that the fine mapping successfully reassigned higher PIP for functional SNPs. We note that credible SNPs with PIP > 0.95 covered only 15.3% of fine-mapped loci, which might bias our observation. We, therefore, further evaluated credible SNPs at PIP > 0.8, 0.5 and 0.1, and showed similar enrichment patterns but with increasing proportion of functional SNPs with increasing PIP threshold (Supplementary Note and Supplementary Table 22).

Nature of the genetic architecture of complex traits. To investigate how the genetic architecture varies across multiple complex traits, we assessed h^2_{SNP} and the polygenicity of 558 traits.

SNP heritability. The h^2_{SNP} is an indication of the total amount of phenotypic variance that is captured by the additive effects of all variants included in a GWAS. The h^2_{SNP} depends on several factors, such as the number of SNPs included in the analyses, the polygenicity of the trait (that is, how many SNPs have an effect) and the distribution of effect sizes. We estimated h^2_{SNP} for each trait using LDSC[17] and SumHer from LDAK[26,27] (see Methods). The estimates of h^2_{SNP} using LDSC and SumHer showed a positive correlation of $r = 0.77$ ($P = 2.5\times10^{-111}$; Fig. 4a). We focus on estimates based on LDSC, hereafter, however, complete results are available in Supplementary Table 23 and are discussed in the Supplementary Note.

The highest h^2_{SNP} was observed for height ($h^2_{SNP} = 0.31$) followed by bone mineral density ($h^2_{SNP} = 0.27$). Of 558 traits, 213 traits, with an average sample size 252,934, showed h^2_{SNP} less than 0.05. Most of these traits are classically regarded as 'environmental' (for example, current employment status, illness of family members or activity regarding lifestyle), and tend to have a low broad sense heritability[14]. For these traits, the number of detected trait-associated loci was also very low, with a median 3. The combination of $N > 200{,}000$ and low h^2_{SNP} suggests that for these traits increasing the sample size may not lead to a substantial increase in detected loci.

Polygenicity and discoverability of complex traits. The general observation from GWASs is that with increasing sample size, detected signals become not only more reliable, but also more numerous, as, with increasing statistical power, smaller SNP effects may be detected. The total number of associated SNPs, the amount of variance they collectively represent, the distribution of effect sizes across the associated SNPs and how many additional individuals are expected to be needed for the detection of a fixed number of novel SNPs are indicators of the polygenicity of a trait. Such polygenicity may vary across traits, and can be informative for designing SNP-discovery studies.

To obtain an indication of trait polygenicity, we applied the causal mixture model for GWAS summary statistics (univariate MiXeR based on a Gaussian mixture model)[28,29] to estimate π (fraction of independent causal SNPs reflecting polygenicity of a trait) and σ_β^2 (variance of effect sizes of the causal SNPs reflecting discoverability of a trait; see Methods). The value of π ranges between 0 and 1, and a high π indicates a high level of polygenicity, while a high σ_β^2 indicates a high level of discoverability of causal SNPs for the trait. Since the standard error of the model estimates become larger for traits with very small h^2_{SNP} due to the small effect sizes, we

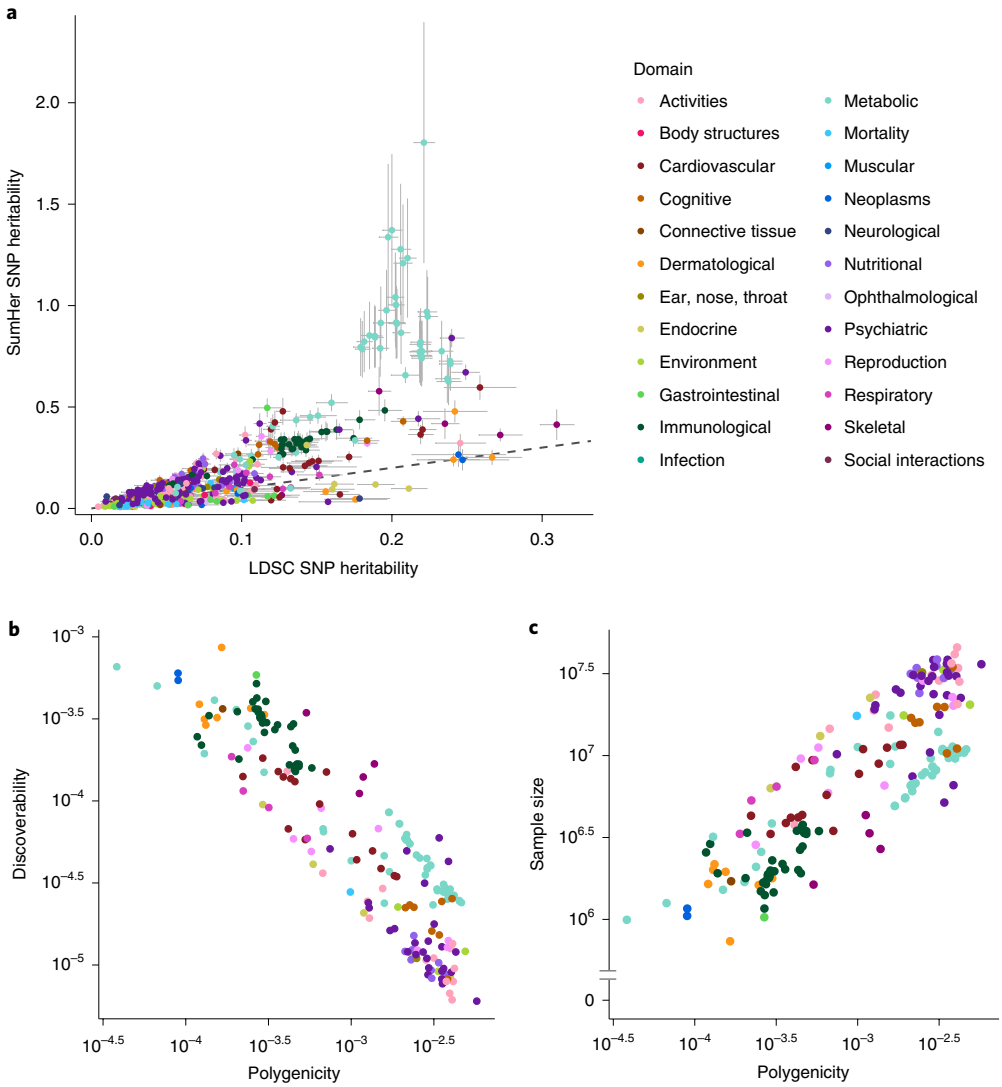

Fig. 4 | SNP heritability and polygenicity of 558 traits. a, Comparison of SNP heritability estimated by LDSC (*x* axis) and SumHer (*y* axis). Horizontal and vertical error bars represent standard errors of LDSC and SumHer estimates, respectively. Full results are available in Supplementary Table 23. **b**, Polygenicity and discoverability of traits, both on $\log_{10}$ scale. Out of 558 traits, 198 traits with reliable estimates (that is, $h^2_{SNP} > 0.05$ (estimated by MiXeR) and standard error of π is less than 50% of the estimated value) are displayed. **c**, Polygenicity and estimated sample size required to reach 90% of total SNP heritability explained by genome-wide significant SNPs, both in $\log_{10}$ scale. Full results are available in Supplementary Table 24. Traits are colored by domain.

only discuss the results of 198 out of 558 traits with $h^2_{SNP} > 0.05$ and standard error of π less than 50% of the estimated value (full results in Supplementary Table 24). We observed, as expected, a negative relationship between polygenicity and discoverability ($r = -0.89$ and $P = 2.4 \times 10^{-70}$), confirming that highly polygenic traits tend to have less causal SNPs with larger effect sizes (Fig. 4b). The majority of traits (that is, 116 of 198 traits) showed high polygenicity with $\pi > 1 \times 10^{-3}$ (more than 0.1% of analyzed SNPs are causal). The highest polygenicity was observed in Major Depressive Disorder, with 0.6% of SNPs being causal, while some traits, such as fasting glucose and serum urate level, showed relatively low polygenicity (Fig. 4b and Supplementary Table 24). The traits with polygenicity >0.1% showed, on average, eight times less discoverability compared with other traits with <0.1% of causal SNPs. The GWAS discoveries for traits with lower polygenicity and high discoverability will saturate with a lower sample size, compared with the traits with higher polygenicity. Indeed, the estimated sample size required to explain 90% of h^2_{SNP} by genome-wide significant SNPs, is positively correlated with polygenicity ($r = 0.83$ and $P = 1.0 \times 10^{-52}$), and extremely polygenic traits require tens of millions of subjects to identify 90% of causal SNPs at a genome-wide significant level (Fig. 4c).

We do note, however, that the model used in the univariate MiXeR assumes both causal and noncausal variants follow Gaussian distributions, $N(0, \sigma_\beta^2)$ and $N(0, 0)$, respectively, which may not hold true for all traits. Therefore, the results should be interpreted as conditional based on these assumptions.

Discussion

Here, we compiled a catalog of 4,155 GWASs to gain insight into the genetic architecture of human complex traits. Based on 558 well-powered GWASs, we addressed fundamental questions concerning the extent of pleiotropy of loci, genes, SNPs and gene sets, characteristics of trait-associated variants and the polygenicity of traits.

We found that the total summed length of trait-associated loci for the 558 analyzed traits covered more than half (60.1%) of the genome. Ninety percent of the grouped loci contained associations with multiple traits across multiple trait domains. High locus pleiotropy can occur in two scenarios: (1) when the same gene in a locus

is associated with multiple traits or (2) when different genes or SNPs in the same locus are associated with multiple traits but, due to LD, the same locus is indicated. Our results showed that locus pleiotropy is widespread (90%), whereas pleiotropy at the level of genes (63%) and SNPs (31%) is much less abundant. This suggests that a gene can be involved in multiple traits but how that gene is affected by the causal SNPs may differ across traits. For instance, the function of the gene can be disrupted through a coding SNP for one trait, while expression of that gene can be affected through a regulatory SNP for another trait. At the same time, overlap of trait-associated loci can be observed due to the overlap of the LD blocks while each trait may be affected by the distinct genes.

Genes and SNPs showing higher levels of pleiotropy were less tissue specific in terms of gene expression and active eQTLs. This suggests that SNPs and genes associated with multiple trait domains are more likely to be involved in general biological functions. Indeed, the most pleiotropic gene sets were mostly involved in regulation of transcription, which is an essential biological function for any kind of cell. Highly pleiotropic genes, therefore, can explain general vulnerability to a wide variety of traits, yet they may be less informative when the aim is to understand the causes of a specific trait. Although a large proportion of trait-associated genes are pleiotropic, the majority of trait-associated gene sets were trait specific. Thus, the trait-specific combination of genes is highly informative, and future studies aimed at improved annotation of gene functions will be needed to understand trait-specific gene-association patterns.

It has been widely acknowledged that almost 90% of GWAS findings fall into noncoding regions[2]. Indeed, our results show that 89.1% of the lead SNPs are noncoding, including intergenic (34.3%) and intronic (43.6%) SNPs. Similarly, of the credible SNPs 89.1% were noncoding (intergenic 30.8% and intronic 47.0%). However, we observed different patterns between intergenic and intronic SNPs; intergenic SNPs were depleted and the intronic SNPs were enriched in both the lead and credible SNPs. We also observed strong enrichment of the lead and credible SNPs in coding and flanking regions. These results indicate that both SNPs with the largest effect size (the lead SNPs) and the most likely causal SNPs (credible SNPs) within a locus tend to be located within or close to the genes.

Our analyses showed that the majority of analyzed traits are highly polygenic, with more than 0.1% of SNPs being causal. For those highly polygenic traits, over tens of millions of individuals are required to identify all SNPs at genome-wide significance ($P < 5 \times 10^{-8}$) that can explain at least 90% of the additive genetic variance (assuming that the distribution of remaining effects follows a Gaussian distribution). In the case of polygenic traits, individuals have almost unique combinations of risk/effect alleles for a specific disease or trait. With higher levels of polygenicity, and thus larger quantities of causal SNPs, the possible combinations of these increase exponentially. This substantially increases the degree of genetic heterogeneity of traits, and complicates the detection of genetic effects, as the effect sizes of individual SNPs that are yet to be detected are even smaller than those observed in current GWASs.

In conclusion, we provide the most comprehensive overview so far, of the extent of pleiotropy, the nature of associated genetic regions and variation in genetic architecture across traits. This knowledge can guide the design of future genetic studies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0481-0.

Received: 24 April 2019; Accepted: 11 July 2019;
Published online: 19 August 2019

References

1. Edwards, A. O. et al. Complement factor H polymorphism and age-related macular degeneration. *Science* **308**, 421–425 (2005).
2. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
3. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011).
4. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
5. Henderson, P. & Stevens, C. The role of autophagy in Crohn's disease. *Cells* **1**, 492–519 (2012).
6. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
7. Gaulton, K. J. et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
8. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
9. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110–124 (2018).
10. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
11. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common disease is more complex than implied by the core gene omnigenic model. *Cell* **173**, 1573–1580 (2018).
12. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
13. Goh, K. et al. The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690 (2007).
14. Polderman, T. J. C. et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).
15. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
16. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
17. Bulik-sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
18. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).
19. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
20. The GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
21. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
22. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
23. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
24. van de Bunt, M., Cortes, A., Brown, M. A., Morris, A. P. & McCarthy, M. I. Evaluating the performance of fine-mapping strategies at common variant GWAS loci. *PLoS Genet.* **11**, e1005535 (2015).
25. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
26. Speed, D. et al. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
27. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2019).
28. Holland, D. et al. Beyond SNP heritability: polygenicity and discoverability estimated for multiple phenotypes with a univariate gaussian mixture model. Preprint at https://doi.org/10.1101/133132 (2018).
29. Frei, O. et al. Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. *Nat. Commun.* **10**, 2417 (2019).

Acknowledgements

We thank all consortiums and all other individual laboratories for making GWAS summary statistics publicly available. We also thank P. Visscher and N. Wray for their thoughtful suggestions and discussions. We additionally thank A. Dale for his suggestions. This work was funded by the Netherlands Organization for Scientific Research (grant nos. NWO VICI 453-14-005 and NWO VIDI 452-12-014).

Author contributions

D.P. designed the study. K.W. curated the database and performed analyses. T.J.C.P. assisted with harmonization of phenotype labels of the database. S.S. performed quality

control on the UK Biobank data and wrote the analysis pipeline for UKB analyses. M.U.M. assisted with the fine-mapping analyses. C.d.L. assisted with the discussion of SNP heritability estimates with different models. O.F. and O.A.A. developed software the MiXeR and assisted with the analyses. S.v.d.S. and B.M.N. discussed and provided valuable suggestions for analyses. K.W. and D.P. wrote the paper. All authors critically reviewed the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0481-0.

Correspondence and requests for materials should be addressed to D.P.

Methods

GWAS summary statistics and preprocessing. Publicly available GWAS summary statistics were curated from multiple resources and were included only when the full set of SNPs were available (last update 23 October 2018). Details are provided in Supplementary Note and Supplementary Table 25.

Additional to the summary statistics available from external studies, we performed GWASs of traits from UK Biobank release 2 cohort (UKB2)[12] with PLINK v.1.9 (ref. [30]) under application ID 16404 (last update 31 October 2017; see Supplementary Note for details). The complete list of GWAS summary statistics is available in Supplementary Table 3.

Definition of lead SNPs and trait-associated loci. For each GWAS, we defined lead SNPs and genomic trait-associated loci as described previously[31]. First, we defined independent significant SNPs with $P < 5 \times 10^{-8}$ and independent at $r^2 < 0.6$, and their LD blocks based on SNPs with $P < 0.05$. Of these SNPs, we further defined lead SNPs that are independent at $r^2 < 0.1$. We finally defined genomic trait-associated loci by merging LD blocks closer than 250 kilobases (kb). Each trait-associated locus was then represented by the top SNP (with the minimum P value) and its genomic region was defined by the minimum and maximum position of SNPs that are in LD ($r^2 \geq 0.6$) with one of the independent significant SNPs within the (merged) locus. We used 1000 Genome Project phase 3 (1000G)[32] as a reference panel to compute LD for most of the GWASs in the database. For each GWAS, the matched population (from African (AFR), American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS)) was used as the reference based on the information obtained from the original study. For trans-ethnic GWASs, the population with the largest total sample size was used. For GWAS based on the UK Biobank release 1 cohort (UKB1), we used 10,000 randomly sampled unrelated white British subjects from UKB1 as a reference. For GWASs based on the UKB2, 10,000 randomly selected unrelated EUR subjects were used as a reference. Multi-allelic SNPs were excluded from any analyses. The reference panel for each GWAS is specified in Supplementary Table 3.

In the current study, lead SNPs with minor allele count (MAC) ≤ 100 (based on MAF and sample size of the SNP) were excluded from any of the analyses, because of lower statistical power and a high false-positive rate among SNPs with extremely small MAF.

MAGMA gene and gene-set analysis. We performed MAGMA v.1.06 (ref. [16]) (https://ctg.cncr.nl/software/magma) gene and gene-set analyses for every GWAS in the database. For gene analysis, 20,260 protein-coding genes were obtained using the R package BioMart (Ensembl build v.92 GCRh37). SNPs were assigned to genes with a 1-kb window both sides. The reference panel was based on either 1000G, UKB1 or UKB2 as described in the previous section. Gene analysis was performed with default parameters (snp-wise mean model). Gene-set analysis was performed for 4,737 curated gene sets (C2) and 5,917 gene ontology terms (C5; 4,436 biological processes, 580 cellular components and 901 molecular functions) from MSigDB v.6.1 (http://software.broadinstitute.org/gsea/msigdb)[33].

SNP heritability and genetic correlation with LDSC. We performed LDSC (https://github.com/bulik/ldsc)[17] for GWASs in the database to estimate SNP heritability and pairwise genetic correlations. Precalculated LD scores for 1000G EUR and EAS populations were obtained from https://data.broadinstitute.org/alkesgroup/LDSCORE/, and LDSC was only performed for GWASs based on either an EUR or EAS population, and when the number of SNPs in the summary statistics file was >450,000. SNPs were limited to HapMap3 SNPs and the MHC region (25–34 Mb) was excluded. When the signed effect size or odds ratio was not available in the summary statistics file, '--a1-inc' flag was used. As recommended previously[34], we excluded SNPs with $\chi^2 > 80$. For binary traits, the population prevalence was curated from the literature (only for diseases whose prevalence was available, Supplementary Table 26) to compute SNP heritability at the liability scale. For most of the (binary) personality/activity traits from the UKB2 cohort, we assumed that the sample prevalence is equal to the population prevalence, since the UK Biobank is a population cohort and not designed to study a certain disease or trait. Likewise, when population prevalence was not available, sample prevalence was used as population prevalence for all other binary traits. Genetic correlations were computed for pairwise GWASs with the following criteria as suggested previously[34]:

- GWASs of EUR population or more than 80% of samples are EUR.
- Number of SNPs is >450,000.
- Signed statistics are available.
- Effect and noneffect alleles are explicitly mentioned in the header or elsewhere.
- Z-score for h^2_{SNP} is >2.

In total, pairwise genetic correlations were computed for 1,193 GWASs in the database.

Selection of GWASs for cross-phenotype analyses. From the 4,155 curated GWASs, we selected 558 GWASs with unique traits for cross-phenotype analyses based on the following criteria.

- N > 50,000 and both cases and controls are >10,000 for binary phenotypes.
- Number of SNPs is >450,000.
- GWAS is based on EUR population or >80% of the samples are EUR. If summary statistics of both trans-ethnic and EUR-only are available, use EUR-only GWAS.
- Exclude sex-specific GWAS, unless the trait under study is only available for a specific sex (for example, age at menopause). If sex-specific and sex-combined GWASs are available, use sex-combined GWAS.
- Z-score of h^2_{SNP} is >2.
- Signed statistics (beta or odds ratio) is available.
- Effect and noneffect alleles are explicitly mentioned in the header or elsewhere.

From GWASs that met the above criteria, for each trait, GWAS with the maximum sample size was selected. UKB2 GWASs performed in this study were further filtered based on the following:

- Exclude cancer screening or test phenotypes.
- Exclude item-level phenotypes (that is, neuroticism and fluid intelligence tests).
- Exclude phenotypes of parents age and parents still alive.
- Exclude medication, treatment, supplements and vitamin traits.
- If exactly the same traits were diagnosed by an expert (for example, doctor) and self-reported, use the expert qualification.
- If exactly the same traits were present as main and secondary diagnoses, include both.
- Phenotypes with large extremes were excluded from the analyses when the difference between the maximum value and 99 percentiles of the standardized phenotype value is >50.

There was one exception, for height GWAS, where a meta-analysis by Yengo et al.[35] (ID 4044) has the largest sample size, however, the meta-analysis was limited to ~2.4 million HapMap2 SNPs. Since over 10 million SNPs are included in most of the selected GWASs, this smaller number of SNPs can bias our analyses. Therefore, the second-largest GWAS (ID 3187) was used instead. This resulted in a total of 558 GWASs across 24 domains. Out of the 558 GWASs, 479 (85.8%) were based on the UKB2, including 11 meta-analyses with UKB2, 46 (8.2%) on the UK Biobank release 1 cohort (UKB1), including 8 meta-analyses with UKB1, and the remaining were non-UKB cohorts. These 558 GWASs are specified in Supplementary Table 3.

Pleiotropic trait-associated loci. To define pleiotropic loci for the 558 GWASs, we first extracted trait-associated loci on autosomal chromosomes. We excluded any locus with a single SNP (no other SNPs have $r^2 > 0.6$) as these loci are more likely to be false positives. Physically overlapping loci were then grouped across 558 traits. In a group of loci, it is not required that all individual trait-associated loci are physically overlapping, but merging them should result in a continuous genomic region. For example, when loci A and B physically overlap and loci B and C also physically overlap, but A and C do not, these three loci were grouped into a single group of loci (Supplementary Fig. 3). Therefore, a grouped locus could contain more than one independent locus from a single trait when gaps between them were filled by loci from other traits. The grouped loci were further assigned to three categories, (1) multidomain locus when a loci group contained traits from more than one domain, (2) domain-specific locus when a loci group contained more than one trait from the same domain and (3) trait-specific locus when a locus did not overlap with any other loci.

We compared the distribution of gene density across three association categories of the loci and nonassociated genomic regions. To define nonassociated genomic regions, we extracted the minimum and maximum positions that were covered by 1000G, and the gap regions of grouped trait-associated loci were defined as nonassociated regions. The gene density was computed as a proportion of a region that was overlapping with one of 20,260 protein-coding genes obtained from Ensembl v.92 GRCh37. We then performed a pairwise Mann–Whitney U-test (two-sided).

Colocalization of trait-associated loci. To evaluate if physically overlapping trait-associated loci also share the same causal SNPs, we performed colocalization using the coloc.abf (approximate Bayes factor colocalization analysis) function from the coloc package in R[36]. Colocalization analysis was performed for all possible pairs of physically overlapping trait-associated loci across 558 traits. When two loci from different traits were physically overlapping but there were no SNPs that were present in both GWAS summary statistics in that overlapping region, colocalization was not performed. The inputs of the coloc.abf function were P value, MAF and sample size for each SNP. When MAF was not available in the original summary statistics, it was extracted from the matched reference panel. For binary traits, sample prevalence was additionally provided based on total cases and controls of the study.

We defined a pair of loci as colocalized when the posterior probability of sharing the same (single) causal SNPs between two traits is >0.9 (Supplementary Note). We note that it is possible that genomic regions outside the predefined trait-associated loci can also colocalize with other traits. However, we limited the analyses to the predefined trait-associated loci in this study, to be consistent with the level of pleiotropy measured by physical overlap of the loci.

Within a grouped locus defined based on physical overlap (see previous section), we further grouped loci based on a colocalization pattern. To do so, we considered the colocalization pattern across a group of physically overlapping loci as a graph in which nodes represent trait-associated loci and edges represent colocalization between loci. First, loci that did not colocalize with any other loci were considered as independent loci. For the remaining loci, we identified connected components of the graph (Supplementary Fig. 3). This does not require all loci within a component to be colocalized with each other. For example, when locus A is colocalized with locus B, and locus B is colocalized with locus C, but locus A is not colocalized with locus C, all loci A, B and C are grouped into a single connected component. Detailed results are discussed in the Supplementary Note.

Pleiotropic genes. For gene-level pleiotropy, we extracted MAGMA gene analysis results for the 558 traits, where 17,444 genes on autosomal chromosomes were tested in all GWASs. For each trait, genes with $P < 2.87 \times 10^{-6}$ (0.05/17,444) were considered as significantly associated. We did not correct the P value for testing 558 traits, since our purpose is not to identify genes associated with one of the 558 traits, but to evaluate the overlap of trait associations (when GWAS was performed for a single trait) across the 558 traits, and this applies to SNPs and gene-set level pleiotropy as well. The trait-associated genes were further categorized into three groups in a similar way as for trait-associated loci, that is, (1) multidomain genes that were significantly associated with traits from more than one domain, (2) domain-specific genes that were significantly associated with more than one trait from the same domain and (3) trait-specific genes that were significantly associated with a single trait. We note that some of the observed gene pleiotropy can still be induced by LD, for example, genes located close to the actual causal gene also tend to show significant association based on the MAGMA gene-based test.

We compared gene length and pLI score across genes in three different association categories and nonassociated genes. Gene length was based on the start and end position of genes extracted from the R package biomaRt, and pLI score was obtained from ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/functional_gene_constraint. We performed t-tests for gene length in log scale and Mann–Whitney U-tests for pLI scores (both two-sided).

For each protein-coding gene, we assessed whether a gene is expressed or not in each of 53 tissue types based on expression profile obtained from GTEx v.7 (ref. [20]) (https://www.gtexportal.org/home/). We defined genes as expressed in a given tissue type if the average transcripts per kilobase million is >1. We then counted the number of tissue types where the gene is expressed and grouped them into six categories, that is, genes expressed in (1) a single tissue type (tissue-specific genes), (2) between 2 and 13, (3) between 14 and 26, (4) between 27 and 39, (5) between 40 and 52 and (6) 53 (all) tissue types. At each number of associated domains (from 1, to 10 or more domains), we recalculated the proportion of genes in each of the six categories, and performed the Fisher's exact tests (one-sided) against all other genes to evaluate if the proportion is higher than expected.

Pleiotropic SNPs. We extracted 1,740,179 SNPs that were present in all 558 GWASs. To evaluate if the selection of ~1.7 million SNPs biased the results, we compared distribution of these analyzed SNPs with all the known SNPs in the genome (SNPs present in 1000G EUR, UKB1 and UKB2 reference panels) by computing the proportion of SNPs per chromosome. In addition, distribution of functional consequences of SNPs annotated by ANNOVAR[37] (http://annovar.openbioinformatics.org/) was compared with all SNPs in the genome.

For each SNP, we counted the number of traits to which an SNP was significantly associated at $P < 5 \times 10^{-8}$, and then grouped the associated SNPs into multidomain, domain-specific and trait-specific SNPs using the same definitions as at the gene level. We note that some of the observed SNP pleiotropy may still be induced by LD, for example, an SNP could reach genome-wide significance because of its strong LD with a causal SNP. However, the purpose of this analysis is to identify individual SNPs (not loci) that are associated with multiple trait domains and their functions.

Functional consequences of SNPs were annotated using ANNOVAR[37]. To test if an SNP from a certain functional category is enriched at a given number of associated domains compared to all analyzed SNPs, a baseline proportion was calculated from the 1,740,179 SNPs for each functional category. At each number of associated domains (from 1, to 10 or more domains), we recalculated the proportion of SNPs with each functional category and performed the Fisher's exact test (one-sided) against the baseline (the proportion relative to all 1,740,179 SNPs), to test if the proportion is higher than expected.

eQTLs for 48 tissue types were obtained from GTEx v.7 (ref. [20]) (https://www.gtexportal.org/home/) and we considered SNPs with gene $q < 0.05$ with any gene in any tissue as eQTLs. For each eQTL, we counted the number of tissue types being eQTL (regardless of associated genes) and categorized them into five groups, that is, being eQTLs in (1) a single tissue type (tissue-specific eQTLs), (2) between 2 and 12, (3) between 13 and 24, (4) between 25 and 36 and (5) being in more than 37 tissue types. At each number of associated domains, we recalculated the proportion of SNPs in each of the five categories, and performed the Fisher's exact test (one-sided) against baseline (the proportion relative to all 1,740,179 SNPs), to test if the proportion is higher than expected.

Pleiotropic gene sets. For gene set-level pleiotropy, we extracted 10,650 gene sets (with at least ten genes) tested in all 588 traits. We then considered gene sets with $P < 4.69 \times 10^{-6}$ (0.05/10,650) as significantly associated. The trait-associated gene sets were grouped into multidomain, domain-specific and trait-specific gene sets with the same definitions as at the gene level.

We compared the number of genes in different association categories and nonassociated gene sets, by performing two-sided t-tests in the log scale of the number of genes.

Power calculation of genetic correlation. Power calculations were performed using the bivariate analysis of GCTA-GRML power calculator (http://cnsgenomics.com/shiny/gctaPower/)[38], to estimate the minimum r_g that obtained a power of 0.8 in the worst-case scenario. From 558 traits, two traits with worst-case scenarios were selected, one with the minimum h^2_{SNP} estimated by LDSC and another with the minimum sample size. For each case, we obtained the minimum r_g to obtain a power of 0.8 by assuming both traits are quantitative with same sample size and h^2_{SNP} and have phenotypic correlation of 0.1.

Hierarchical clustering of trait based on genetic correlation. Hierarchical clustering was performed on the matrix of pairwise r_g values as calculated between the 558 traits. After Bonferroni correction for all possible trait pairs, nonsignificant genetic correlations were replaced with 0. The number of clusters, k, was optimized between 50 and 250 by maximizing the silhouette score with 30 iterations for each k.

Estimated standardized effect size of lead SNPs. To enable comparison of effect sizes across studies, we first converted P values into Z-statistics (two-sided) and expressed the estimated standardized effect size (β) as a function of MAF and sample size as described previously[21] using the following equation:

$$\beta = \frac{z}{\sqrt{2p(1-p)(n+z^2)}}, \text{ s.e.m.} = \frac{1}{\sqrt{2p(1-p)(n+z^2)}}$$

where p is MAF and n is the total sample size. We used the MAF of a corresponding European reference panel (either 1000G, UKB1 or UKB2) as described in the section 'Definition of lead SNPs and genomic trait-associated loci'. Since we were not interested in the direction of effect, we used squared standardized effect sizes for analyses in this study.

Fine mapping of trait-associated loci. We defined the region to fine map by taking 50 kb around the top SNPs of the trait-associated loci. When trait-associated loci were larger than the 50-kb window, the largest boundary was taken. Due to the complex LD structure, loci overlapping with the MHC region (chr6: 25–36 Mb) were excluded. The fine mapping was performed using FINEMAP with the shotgun stochastic search algorithm[25] (http://www.christianbenner.com/#). We used randomly selected 100,000 unrelated European individuals from the UKB2 cohort as a reference panel to estimated pairwise LD correlation of SNPs using LDstore[39] (http://www.christianbenner.com/#) for all 558 GWASs. We limited the number of maximum causal SNPs (k) per locus to 10. When the number of SNPs within a locus is relatively small (around 30 or less), the algorithm can fail to converge. In that case, k was decreased by 1 until FINEMAP was run successfully. Loci with less than 10 SNPs were excluded from the fine mapping. To select most likely causal SNPs, we defined credible SNPs as SNPs with PIP > 0.95. Detailed results are discussed in the Supplementary Note.

Annotation and characterization of lead SNPs and credible SNPs. Functional consequences of SNPs were annotated using ANNOVAR[37] (http://annovar.openbioinformatics.org/) based on Ensembl gene annotations on hg19. Before ANNOVAR, we aligned the ancestral allele with dbSNP build 146. Core 15-state chromatin states of 127 cell or tissue types were obtained from Roadmap[40] (http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/all.mnemonics.bedFiles.tgz) and we annotated one of the 15-core states to each of the lead SNPs based on chromosome coordinates. Subsequently, the consequence state was assigned by taking the most common state across 127 cell or tissue types. SNPs with consequence state ≤7 were considered as active. eQTLs in 48 tissue types were obtained from GTEx v.7 (ref. [20]) (https://www.gtexportal.org/home/) and we only used the significant eQTLs at gene $q < 0.05$. eQTLs were assigned to SNPs by matching chromosome coordinate and alleles.

As we showed that trait-associated loci have higher gene density compared to nonassociated regions, and GWAS signals are known to be enriched in regulatory elements[41], we first identified background enrichment by comparing SNPs within trait-associated loci or fine-mapped regions with the entire genome. For this, all known SNPs were extracted by combining all SNPs in 1000G EUR, UKB1 and UKB2 reference panels (~28 million SNPs in total). SNPs within the trait-associated loci were defined as the ones with $P < 0.05$ and $r^2 > 0.6$ with one of the independent significant SNPs as described above (see section 'Definition of lead SNPs and trait-associated loci'). Therefore, it was not necessary to include all SNPs physically located within the trait-associated loci. On the other hand, SNPs within the fine-mapped region include all SNPs physically located within a 50-kb window

from the most significant SNP of a locus. To characterize lead SNPs and credible SNPs given background enrichments, we compared these SNPs against all SNPs within trait-associated loci or fine-mapped regions, respectively. We performed two-sided Fisher's exact test for each category of annotations.

SNP heritability estimation with SumHer using the LDAK model. We estimated SNP heritability of 558 traits using the SumHer function from LDAK v.5.0 (ref. [27]) (http://dougspeed.com/ldak/). Since our purpose was to compare estimates from LDSC and SumHer, we used the 1000G EUR reference panel and extracted HapMap3 SNPs as consistent with LDSC. We used unique IDs of SNPs (consisting of chromosome, position and alleles) instead of rs ID to maximize the match between GWAS summary statistics and the reference panel. The MHC region (chr6: 25–34 Mb) was excluded. To be consistent with LDSC, SNPs with $\chi^2 > 80$ were excluded. As recommended by the author, SNPs with in LD ($r^2 > 0.1$) with one of those SNPs with $\chi^2 > 80$ were additionally excluded.

To obtain SNP heritability on a liability scale, we provided population prevalence and sample prevalence for binary traits. The same population prevalences were used as described in the section 'SNP heritability and genetic correlation with LDSC' (Supplementary Table 26). Details of the results are discussed in Supplementary Note.

Estimation of polygenicity and discoverability with MiXeR. In the causal mixture model for GWAS summary statistics (univariate MiXeR, https://github.com/precimed/mixer) proposed by Holland et al., the distribution of SNP effect sizes is treated as a mixture of two distributions for causal and noncausal SNPs, as follows[28]:

$$\beta = \pi N\left(0, \sigma_\beta^2\right) + (1-\pi)N(0,0)$$

where π is the proportion of (independent) causal SNPs and σ_β^2 is the variance of the effect sizes of causal SNPs. Therefore, π and σ_β^2 respectively represent polygenicity and discoverability of the trait. We estimated both parameters for the 558 traits using MiXeR software[28,29]. As recommended in the original study, we used 1000G EUR as a reference panel and restricted to HapMap3 SNPs. SNPs with $\chi^2 > 80$ and the MHC region (chr6: 26–34 Mb) were excluded. To estimate the sample size required to explain 90% of the additive genetic variance of a phenotype, we used an output of GWAS power estimates calculated in the MiXeR software, which contains 51 data points of sample size and the proportion of chip heritability explained[28]. We then estimated the sample size required to reaches 90% by using the interp1 function from the pracma package in R.

Statistical analysis. When a Mann–Whitney *U*-test was performed, the null hypothesis was that the distributions of two sets of observations were equal. When a *t*-test was performed, the null hypothesis was that the averages of two sets of observations were equal. When a Fisher's exact test was performed, the null hypothesis was that the having a specific annotation was equally likely across two categories. We report two-sided *P* values of each statistical test unless otherwise specified.

Reporting Summary. Further information on research design is available in the Life Sciences Reporting Summary linked to this article.

Data availability

All publicly available GWAS summary statistics (original) files curated in this study are accessible from the original links provided at https://atlas.ctglab.nl. GWAS summary statistics for 600 traits from UK Biobank performed in this study are also provided at https://atlas.ctglab.nl and an archived file will be made available upon publication from https://ctg.cncr.nl/software/summary_statistics.

References

30. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
31. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
32. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
33. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
34. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
35. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
36. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
37. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
38. Visscher, P. M. et al. Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet.* **10**, e1004269 (2014).
39. Benner, C. et al. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017).
40. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
41. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin* **8**, 57 (2015).

nature research

Corresponding author(s): Danielle Posthuma

Last updated by author(s): Jun 27, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used published summary statistics or performend GWAS based on data collected by UKBiobank
Data analysis	GWASs of UKB cohort was performed with PLINK v1.9, gene and gene-set analyses were performed with MAGMA v1.06, lead SNPs and trait-associated loci for each GWAS were defined using FUMA v1.3.3, colocalization was performed with coloc package in R, fine-mapping was performed with FINEMAP v1.3, functional consequence of SNPs were annotated with ANNOVAR, SNP heritability and genetic correlation were estimated using LD score regression and SumHer implemented in LDAK v5, polygenicity and discoverability was estimated using MiXeR, all other statistical analyses were performed with R v3.4.3.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All publicly available GWAS summary statistics (original) files curated in this study are accessible from the original links provided at http://atlas.ctglab.nl. GWAS summary statistics for 600 traits from UK Biobank performed in this study are also provided at http://atlas.ctglab.nl and an archived file will be made available upon publication from https://ctg.cncr.nl/software/summary_statistics.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used maximum number of samples for UKB GWASs with limiting the subjects to unrelated European ancestry and no-missing phenotype and covariates for each trait.
Data exclusions	For main analyses in the manuscript we selected 558 GWASs from 4155 GWASs curated in the ATLAS database, based on sample size. When there were multiple GWASs with the same trait, GWAS with the largest sample size was selected.
Replication	Replication is not applicable: we analyzed all available GWAS results to obtain a global overview of pleiotropy, genetic architectures and genetic correlations.
Randomization	N.A.
Blinding	N.A.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	We utilized data collected previously by UK biobank. All individuals included in the study provided informed consent, and the study was approved by the concerned ethical committee.
Recruitment	See above (and in Methods section of the manuscript)
Ethics oversight	NHS Health Research Authority provided ethics approval for the UKB study

Note that full information on the approval of the study protocol must also be provided in the manuscript.